DOCUMENT RESUME

ED 228 314 TM 830 240

AUTHOR Cason, Gerald J.; And Others
TITLE Controlling Rater Stringency Error in Clinical Performance Rating: Further Validation of a Performance Rating Theory.
PUB DATE Apr 83
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Clinical Experience; *Error of Measurement; Evaluation Methods; Higher Education; Latent Trait Theory; Longitudinal Studies; *Mathematical Models; *Rating Scales; *Student Evaluation; Test Reliability; Test Theory; Test Validity
IDENTIFIERS Generalizability (Tests); *Performance Rating Theory (Cason and Cason); *Rater Reliability; Rater Stringency Error

ABSTRACT

Prior research in a single clinical training setting has shown Cason and Cason's (1981) simplified model of their performance rating theory can improve rating reliability and validity through statistical control of rater stringency error. Here, the model was applied to clinical performance ratings of 14 cohorts (about 250 students and 200 raters) in a Junior year clerkship at a different medical school. The model fit every cohort's data. In the 11 cohorts where a significant rater effect was observed, reliability and validity of ratings were improved by using the model to obtain adjusted ratings from which rater stringency error was removed. Statistical control of stringency error in these cohorts produced adjusted ratings with a mean standard error of about 2.9 percent compared with 8.4 percent for (unadjusted) observed ratings. These improvements can be had at a fraction of the cost for methods such as rater training. The results provided further support for the validity, extended the generalizability, and illustrated the practical value of the Casons' model. Contained in the appendices are: "A Theory of Performance Rating"; and "MERLIN's Magic: The Estimation of RRP's and SAP's." (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED228314

Controlling Rater Stringency Error in Clinical Performance Rating: Further Validation of a Performance Rating Theory*

Gerald J. Cason
Carolyn L. Cason
University of Arkansas for Medical Sciences
and
John H. Littlefield
University of Texas Health Science Center at San Antonio

Abstract

Prior research in a single clinical training setting has shown Cason and Cason's simplified model of their performance rating theory can improve rating reliability and validity through statistical control of rater stringency error. Here, the model was applied to clinical performance ratings of 14 cohorts (about 250 students and 200 raters) in a Junior year clerkship at a different medical school. The model fit ($0.89 \geq R \geq 0.64$; $p < .01$) every cohort's data. In the 11 cohorts where a significant rater effect (each $p < .05$) was observed, reliability and validity of ratings were improved by using the model to obtain adjusted ratings from which rater stringency error was removed. Statistical control of stringency error in these cohorts produced adjusted ratings with a mean standard error of about 2.9 percent compared with 8.4 percent for (unadjusted) observed ratings. These improvements can be had at a fraction of the cost for methods such as rater training. The results provided further support for the validity, extended the generalizability, and illustrated the practical value of the Casons' model.

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. J. Cason

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 830 240

*Presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada, April 1983.

Address correspondence to:

Gerald J. Cason
Office of Educational Development, Slot 595
University of Arkansas for Medical Sciences
4301 West Markham
Little Rock, AR 72205
Phone: (501) 661-5720

Controlling Rater Stringency Error in Clinical Performance Rating: Further Validation of a Performance Rating Theory*

Gerald J. Cason
Carolyn L. Cason
University of Arkansas for Medical Sciences
and
John H. Littlefield
University of Texas Health Science Center at San Antonio

Objective

This study provided an additional test of Cason and Cason's (1981) quantitative model of performance rating using data sufficiently different from that used in previous research (e.g., different locale; shorter inventory; categorical rather than norm-referenced rating scale) to clarify the model's generalizabilty across settings where statistical control of rater stringency error may be useful because direct control is ineffective or impractical.

Theory

Cason and Cason's (1981) simplified model of their performance rating theory accounts for systematic variation in performance ratings exclusively by variation in rater stringency and student ability. See Appendix A for a detailed exposition of the theory and simplified model. Cason and Cason (1981, 1982; Cason, Cason, & Littlefield, 1983) have presented evidence that this model fits clinical performance rating data in a commom type of setting, i.e., where each student is rated by several but not all the raters and each rater rated several but not all the students. Further, where there were sufficient overlaps in who rated whom, variation in rater leniency was statistically controlled. This substantially improved the reliability and validity of the corrected clinical performance ratings. Although the initial study provided strong validation for the model, it was limited in several important respects.

Except for a single set of data from another school (Cason, Cason, & Littlefield, 1983), the Casons' data were all from a medicine clerkship (i.e., clinically oriented Junior year course) in a single medical school. A single 32 item rating instrument was used. It had been in use for only a short time. Little rater training was given. The rating inventory used explicitly norm-referenced definitions for the scale points (e.g.,"5=Student would be in top 10% of typical class"). This might be more compatible with the theory than other methods for defining the response categories. In addition, the original study reported results from only 3 cohorts of students for a total of about 100 students and 75 raters.

*Presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada, April 1983.

Method and Data Source

Cason and Cason's model and the general linear model (Ward & Jennings, 1973) were used to analyze two years (1981-82) of data available from an on-going longitudinal study of clinical performance evaluation in a medicine clerkship at University of Texas Health Science Center at San Antonio. Results of the first 4 years of that study were reported by Littlefield et al (1981). Each year 7 cohorts of about 20-25 students each went through the clerkship. The modal student was rated by 5 out of approximately 200 raters (faculty attending and resident physicians). The locally developed 5 item inventory used categorical rather than norm-referenced definitions, e.g., "0-2: Unsatisfactory performance with severe deficits." The instrument had been in use at least 4 years (1977-80) prior to the two years of data analyzed here. Raters were provided information annually on how their ratings compared with other raters. As a result, the single-rater reliability increased from about 0.2 in 1977 to about 0.4 in 1980 (Littlefield et al, 1981). Although r=0.4 was approximately the same as the initial single-rater reliability observed in Cason and Cason's original study (after application of their model single-rater reliability rose to 0.6), the differences between the two settings provided a good test of the model's generalizability. Appendex B gives the technical details of the numeric and analytic procedures used to determine fit of the Casons' model to the data and rater effects.

Results

Table 1 shows that the model, which permits but does not require a separate contribution to the observed ratings from variations in both student achievement and rater stringency, fit ($0.89 > R > 0.64$; $p<0.01$) every cohorts' data.

Table 1

Fit of Model to Each Cohort's Data

Year: 1981

Cohort	A	B	C	D	E	F	G
Multiple R*	.85	.70	.75	.69	.64	.89	.76
N of Raters	37	28	46	38	29	30	31
N of Students	21	20	29	24	17	18	20
Total Ratings	108	76	126	102	73	74	91

Year: 1982

Cohort	A	B	C	D	E	F	G
Multiple R*	.68	.76	.71	.81	.77	.78	.72
N of Raters	42	42	34	39	43	39	40
N of Students	24	24	16	21	27	25	27
Total Ratings	129	129	80	114	145	131	133

*All R's significant at $p<.01$.

Table 2 shows that a statistically significant ($p<.05$) rater stringency effect was found in 5 of the 7 cohorts during 1981 and 6 of 7 during 1982. It should be noted that failure to find a significant stringency effect in the remaining 3 cohorts does not mean that there was no variation in standards used by these individual raters. Rather, these results indicated that in these cohorts the mean stringency of the group of raters who rated each student was statistically equal to (i.e., not different from) the mean across all raters for that cohort.

Table 2

Presence of Rater Stringency Effect

Year: 1981

Cohort	A	B*	C	D	E*	F	G
F	4.71	1.76	2.61	2.10	1.69	7.19	3.09
df1/df2	36/51	27/29	45/52	37/41	28/28	29/27	30/41
p	<.001	<.070	<.001	<.012	<0.86	<.001	<.001

Year: 1982

Cohort	A	B	C*	D	E	F	G
F	2.84	1.95	1.47	4.33	4.43	2.82	2.15
df1/df2	41/64	41/87	33/31	38/55	42/76	38/68	39/67
p	<.001	<.005	<.142	<.001	<.001	<.001	<.004

*Nonsignificant rater stringency effect ($p>.05$).

Tables 3 and 4 show the impact of systematic rater stringency error upon the reliability of performance ratings and the gains in the reliabilities had by adjusting for variation in rater stringency. Adjusted ratings were obtained by selecting a fixed rater stringency (e.g., mean stringency) and calculating the expected ratings for all students based on their Subject Ability Points (SAP's). Reliabilities for observed ratings were computed as recommended by Ebel (1951) for cases involving missing data (see Ebel's Table 2). This analysis of variance based procedure gives a single rater reliability (intra-class correlation) and the reliability of the harmonic mean (mean k) ratings per subject as recommended by Snedecor (1946). Because different subjects were rated by different numbers of raters, mean k was used in estimating the reliability of a typical subject's mean rating.

Reliabilities for the adjusted ratings, that is those from which systematic rater stringency effect had been removed, were estimated as follows. The rater effect (sum of squares) was estimated by a rater's product sum analogous to the subject product sum shown in Ebel's Table 2. The rater sum of squares was then subtracted from the residual error sum of squares. In this case error degrees of freedom were defined using mean k as the number of raters per subject (see Ebel's analysis of variance source table). While this method of estimating reliability for adjusted ratings was conservative, it assumed that the simplified model of the Casons' theory is true; that is, that the components of variance identified by the simplified model and only those (or ones isomorphic with them) did in fact

exist.

Table 3

Single Rater Reliabilities

Year: 1981

Cohort	A	B	C	D	E	F	G
Observed Ratings	.30	.31	.24	.22	.22	.16	.23
Adjusted Ratings	.87	(.61)*	.80	.87	(.84)*	1.00	.96

Year: 1982

Cohort	A	B	C	D	E	F	G
Observed Ratings	.08	.24	.40	.18	.10	.35	.33
Adjusted Ratings	.37	.65	(.76)*	.89	.58	.76	.88

*Where there was no significant rater effect values are enclosed in parentheses.

The results in Table 3 are consistent with those reported in Table 2. Adjusting ratings to remove rater stringency effect resulted in higher estimated reliability for the adjusted than the observed ratings in all cases. Even in those cohorts where the rater effect did not reach statistical significance, the estimated reliability of adjusted ratings was higher. These cases are indicated in Table 3 by reliability values being enclosed in parentheses. Single rater reliabilities for the observed ratings were systematically lower than those reported by Littlefield et al. (1981) for earlier years. This may have resulted from differences in the procedures used to make the estimates.

Table 4

Reliabilities of Aggregate of k Ratings

Year: 1981

Cohort	A	B	C	D	E	F	G
Mean k	5.13	3.79	4.33	4.24	4.28	4.08	4.53
Observed Ratings	.69	.63	.57	.54	.55	.45	.57
Adjusted Ratings	.97	(.86)*	.95	.97	(.96)*	1.00	.99

Year: 1982

Cohort	A	B	C	D	E	F	G
Mean k	5.37	5.37	4.99	5.42	5.37	5.23	4.92
Observed Ratings	.32	.63	.77	.54	.38	.74	.71
Adjusted Ratings	.76	.91	(.94)*	.98	.88	.94	.97

*Where there was no significant rater effect values are enclosed in parentheseses.

The reliabilities reported in Table 4 indicate the improvment in measurement precision that would have been achieved in these specific groups had adjusted ratings rather than the mean of observed ratings been used for evaluating students. In Table 4 the number of raters rating a typical student (mean k) was used in calculating the reported reliabilities.

The practical gain in measurement accuracy is perhaps most concisely communicated in terms of the standard error of measurement (McNemar, 1966). Table 5 reports the standard error of measurement of aggregate performance rating scores expressed as percent scores. (The possible total rating on the original instrument ranged from 0-70 raw points.) The mean k and reliabilities on which these standard errors are based are given in Table 4 for each group. The standard deviation of the observed was used in the computation for the reliability of both the observed and adjusted ratings.

Table 5

Standard Error for Aggregate of k Ratings

	Year: 1981						
Cohort	A	B	C	D	E	F	G
Observed Ratings	7.74	9.11	8.54	8.41	8.07	10.52	7.72
Adjusted Ratings	2.41	(5.60)*	2.91	2.15	(2.41)*	0.00	1.18
	Year: 1982						
Cohort	A	B	C	D	E	F	G
Observed Ratings	11.80	8.15	6.13	9.41	10.07	6.98	7.14
Adjusted Ratings	7.01	4.02	(3.13)*	1.96	4.43	3.35	2.30

*Where there was no significant rater effect values are in parentheses.

The results reported in Table 5 may be roughly summarized by the difference between the unweighted mean standard errors for observed and adjusted ratings in only those groups having a signficant rater stringency effect. Mean for observed was 8.4 while for adjusted it was 2.9. A 5 percentage point decline in the standard error is certainly both statistically and practically significant.

The improvements in clinical performance measurement accuracy reported above were achieved by removing a theory specific systematic error component from the observed ratings. Because a systematic source of error was controlled, the adjusted ratings were more reliable. They were also more nearly valid because of the construct validity of the model used to obtain the adjusted ratings. The results contribute to further validating the constructs of the theory; while the theory contributes to increased validity of the adjusted ratings.

Conclusions

Cason and Cason's (1981) model has fit performance rating total scores from medicine clerkships at two different schools with different rating inventories and different amounts of rater training and rater familiarity with the respective instruments. The model has been applied to the data of about 375 students rated by 275 raters during a span of 5 years. Overwhelmingly the evidence supports the model. That a significant rater

stringency effect was found in 11 of 14 cohorts in this study indicates that casual rater training and annual feedback to raters may not be sufficient to eliminate variations in rater stringency, even after 4 years. The validity and reliability of clinical performance ratings could have been substantially improved by statistical correction for rater stringency error in all those groups in which there was a signficiant rater stringency effect. Taken together the results support the validity, extend the generalizability, and illustrate the practical value of the Casons' model.

Importance

The reported research indicates that Cason and Cason's theory provides a useful approach to the study and understanding of the performance rating process, as well as, an explicit basis for a practical technology which in some cases will allow rating complex human performance independent of some individual rater characteristics, even when practical constraints prevent using extensive rater training or other direct control techniques. When statistical control afforded by Cason and Cason's model frees ratings-based measures of performance from rater bias (e.g., stringency), substantially more nearly reliable and valid measures result. The improved reliability and validity of such measures, directly improves the professional and educational decisions which may be based upon them. Of equal practical importance, gains had by statistically correcting for rater stringency can be had for a small fraction of the costs associated with direct control measures such as rater training. Furthermore, where training is practical, the methods are complimentary.

References

Cason, G. J., & Cason, C. L. Some promising early results from a rudimentary latent-trait theory of performance rating. Paper presented at the annual meeting of AERA, Los Angeles, 1981. Abstract in Resources in Education, 1981, 16(9), 152. ERIC Document number ED-201-669.

Cason, G. J., & Cason, C. L. Application of latent-trait theory to clinical performance rating. In C. Friedman (Chair), Latent-trait models: How useful are they to professional education. Paper presented in symposium session 34.14 at the annual meeting AERA, New York, 1982. Symposium audio tape R-34.14 (two cassettes) available from: AERA Cassettes, c/o Teach 'em, Inc., 160 E. Illinois St., Chicago, IL 60611.

Cason, C.L., Cason, G.J., & Littlefield, J.H. Variation in intra-rater stringency in cognitive-technical and affective-interpersonal clinical performance domains. Presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada, April 1983.

Ebel, R.L. Estimation of reliability of ratings. Psychometrika, 1951, 61, 407-424.

Littlefield, J. H., Harrington, J. T., Anthracite, N. E., & Garman, R. E. A description and four-year analysis of a clinical clerkship evaluation system. Journal of Medical Education, 1981, 56, 334-340.

McNemar, Q. Psychological statistics. (3rd Ed.) New York: Wiley, 1966.

Snedecor, G.W. Statistical methods. (4th Ed.) Ames, Iowa: Iowa State College Press, 1946.

Ward, J., & Jennings, E. Introduction to linear models. Englewood Cliffs, N.J.: Prentice-Hall, 1973.

A Theory of Performance Rating

Gerald J. Cason and Carolyn L. Cason
University of Arkansas for Medical Sciences

Development of our theory was not directly motivated by recent developments in latent trait test theory (Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1978; Wright & Stone, 1978) nor item characteristic curve analysis (Baker, 1977). Nor was our work a conscious extrapolation of the models of Lord (1952) or Rasch (1966). However, the generic, intuitive concepts underlying Lord's and Rasch's models are very closely related to ours. This similarity can mislead those familiar with current testing theory if exact equivalence of meaning and methods is expected. For the most part, test theory assumes dichotomous response data from examainees to test items. Our theory applies to continuous rating data from judges regarding subjects. Test theory is concerned with characteristics of items and examinees. We are concerned with characteristics of judges and subjects rather than inventory items and survey respondents. Differences in basic data prevent use of most of the technical methods appropriate to test data. Differences in focus produce a real but subtle change in the meaning of some basic concepts.

Our Theory

We propose that the rating received by a subject is a function of the subject's true ability and the rater's characteristics, including the rater's resolving power, sensitivity, stringency, and effective rating floor and ceiling. Rater characteristics define the relationship between subject ability and assigned rating. They are illustrated by the rater characteristic curve (RCC) in Figure 1. The RCC embodies the rater's knowledge, understanding, and beliefs about the subject whose performance is to be rated, the task to be performed and its difficulty, the rating procedure including the rating instrument, external constraints imposed by the setting, and related factors. It results from the rater's effort to take all the discrepancies between an ideal setting and the actual one into account prior to assessing the subject's performance.

Insert Figure 1 about here

Resolving power (RP) is the rater's capacity to assign different ratings to different amounts of subject ability. Resolving power is reflected in the effect of a fixed increment in ability on the increment in assigned rating. Where the increment in ability leads to a large increment in rating, resolving power is high. Resolving power is assumed to have a single maximum and to decline in an accelerated, continuous manner on each side of this maximum such that the RCC is a smooth, s-shaped curve as illustrated in Figure 1. The measure of resolving power is the slope of the RCC.

Rater sensitivity is the maximum value of rater resolving power. Rater sensitivity is positively correlated with the rate of change in resolving power. The measure of sensitivity is the maximum of the RCC slope (e.g., at point M in Figure 1). Sensitivity and therefore rate of change in resolving power may vary between raters.

Rater stringency is the general tendency to require higher or lower ability for a given rating when other characteristics across raters are equal. Both subject ability and rater stringency are measured on the same underlying equal-interval scale. 'Ability' is used in its broadest sense and encompasses innate and acquired capacity, skill, disposition, and so forth. In Figure 1, stringency is depicted by the location of the RCC along the ability and stringency scale. The exact location of the rater on the ability and stringency scale is called the rater reference point (RRP). The RRP is located on the scale at the projection of that point on the RCC at which the rater has maximum resolving power. Thus, Figure 1 illustrates the RRP as the projection of point M (i.e., where RCC slope is maximum) onto the ability and stringency scale at point L. The value of the RRP on the ability and stringency scale is the measure of the rater's stringency.

The RRP may be intuitively understood as the rater's pivotal standard for judging perceived ability. It is as if the rater perceives the subject's performance as a deviation from his rater reference point and as corresponding to a point on the ability and stringency scale. This point is called the subject ability point (SAP). In Figure 1, this point is illustrated by point S. The rater evaluates the subject's performance in terms of the perceived SAP's deviation from his standard (RRP) and other characteristics defining his RCC. In Figure 1, the RCC evaluates the difference between the RRP at L and the SAP at S as rating R.

The effective rating ceiling (ERC) and effective rating floor (ERF) are the maximum and minimum ratings a rater will assign. The effective rating ceiling and effective rating floor subsume the concept of rater central tendency error. The effective rating ceiling and effective rating floor arise from an interaction between the manner in which the rating scale is defined and the rater's knowledge and beliefs about the distribution of subject ability. For example, Figure 1 depicts a rater whose effective rating floor is 30% of the maximum value on the rating scale (i.e., instrument). This could reflect the rater's belief that regardless of apparent performance, below the point corresponding to the rating instrument's definition of 30% (e.g., 3 on a 10 point instrument) performance reflects extraneous factors such as subject illness rather than ability. The measures of effective rating ceiling and effective rating floor are respectively the upper and lower limits of the domain of the RCC on the rating scale. They each may vary from rater to rater.

Insert Figure 2 about here

Holding constant all rater characteristics except stringency, Figure 2 illustrates the effect of variation of stringency on rating assigned. Figure 2 depicts the RCCs of raters A and B with RRPs located at K and L, respectively. Should raters A and B both observe a subject with an SAP located at S, A would assign a rating of RA while B would assign a rating of RB. As the value of an RRP increases on the ability and stringency scale, the rater becomes more stringent, i.e., assigns a lower rating for a given ability.

Insert Figure 3 about here

The impact of rater sensitivity is shown, in part, in Figure 3. Holding constant all characteristics except sensitivity, rater A would assign ratings RA1 and RA2 and rater B would assign ratings RB1 and RB2 to

a pair of subjects with SAPs at S1 and S2. Note that even though both raters have a common RRP (at L), rater A's greater sensitivity (slope at M) results in harsher judgement of S1 and S2 than does Rater B's. This illustrates that the intuitively useful definition of stringency as the rater's general tendency to be more or less demanding is limited to cases where other rater characteristics are equal as shown in Figure 2. Figure 3 also gives a specific example of the effect of sensitivity on resolving power. Rater A's greater sensitivity results in a larger increment in ratings than rater B's. That is, the difference between RA1 and RA2 is larger than the difference between RB1 and RB2 for a fixed increment in ability (S1 to S2) near L. This effect is illustrated more generally in Figure 4 which shows that rater A's high sensitivity is associated with high resolving power over a narrow range of RRP to SAP differences while rater B's lower sensitivity yields a lower resolving power but spread across a broader range of ability. Thus, in general, the highly sensitive rater can provide more information about subjects closely clustered aroung his RRP and the less sensitive rater can provide more information about subjects that are broadly dispersed about his RRP.

Insert Figure 4 about here

In this theoretic view actual rating instruments are not directly linked to that which they are intended to measure. The true measure of the variable of interest is on the ability and stringency scale. The link between any point on the instrument's scale and the ability scale is through a rater's characteristics. Therefore, a particular point on the actual instrument's scale has equivalent meaning only across raters with equal characteristics. Furthermore, there is no necessary equivalence between the midpoint of the instrument's scale and the rater's reference point. For a rater with the characteristics illustrated in Figure 1, zero difference between RRP and SAP evaluates to 65% on the instrument which is certainly not the midpoint on a percent scale.

Our Simplified Performance Rating Model

To facilitate our initial empirical investigations, we imposed the following simplifying assumptions upon the rudimentary theory presented above:

1. All raters have equal sensitivity.
2. The rater characteristic curve evaluates the difference between a rater reference point (RRP) and subject ability point (SAP) as the percent (%) of possible rating points (0 to 100%).
3. The effective rating floor equals 0% and the effective rating ceiling equals 100%.

More formally, we propose that the scale upon which RRP's and SAP's are located is an equal interval ability and stringency scale of arbitrary graduation (unit) and arbitrary origin (zero point). For the purposes of this research, we defined the rater characteristic curve as the product of an arbitrary positive, constant scaling factor (SF) and the cumulative unit-normal deviate ogive. From the above it follows that, the rater reference point (RRP) for any rater is located under that rater's characteristic curve (RCC) on the ability and stringency scale at that point which evaluates to a rating of 50%. SF is arbitrarily set equal to 100. The difference between RRP and SAP divided by the scaling factor

gives an ability and stringency scale deviation value (z):

Formula 1
$$z = (SAP - RRP) / SF$$

The proportion of possible rating points assigned for a given value of z is equal to the proportion of area under the unit-normal curve below z, that is $p(z)$. Multiplying the proportion $p(z)$ by 100 gives the expected subject rating (ESR) in percent units:

Formula 2
$$ESR = p(z) * 100$$

The relationship between the expected subject rating (ESR) and the discrepancy between RRP and SAP is depicted graphically in Figure 5.

Insert Figure 5 about here

There may be variation in the rater's perception, knowledge, judgement, and so forth. Therefore, the observed subject rating (OSR) may contain error:

Formula 3
$$OSR = ESR + error$$

As all rater characteristics except stringency are equal across raters, this model accounts for systematic variation in observed ratings by variations in only rater stringency and subject ability as illustrated in Figure 2.

References

Baker, F.B. Advances in item analysis. Review of Educational Research, 1977, 47, 151-178.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R., & Gifford, J.A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48. 467-510.

Lord, F.M. A theory of test scores. Psychometric Monographs, 1952, No. 7.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Wright, B.D., & Stone, M.H. Best test design. Chicago: MESA Press, 1979.

MERLIN's Magic: The Estimation of RRP's and SAP's

Gerald J. Cason and Carolyn L. Cason
University of Arkansas for Medical Sciences

Although Hambleton, Swaminathan, Cook, Eignor, and Gifford (1978) suggest that the parameter estimation problem for the continuous (rating) case has been solved by Samejima (1973), Samejima's factor analysis related approach appears to be restricted to complete data matrices. Therefore, a special purpose program based on a different approach from Samejima's was developed to estimate the parameters. Program MERLIN (Cason, 1980) was used in conjunction with subroutine STEPIT (Chandler, 1965) to obtain least-squares estimates of the rater reference points and subject ability points. Briefly, MERLIN operates as follows. An observed data table with one row per subject and one column per rater is input. All observed subject ratings (OSR) are contained in this data table. A set of "best guesses" for the RRP's and SAP's are input. The program uses these starting guesses for the SAP's and RRP's and the normal ogive function depicted in Figure 5 to calculate an expected subject rating (ESR) for every cell in an expected data table. Then, the discrepancy between each value in the observed data table and its corresponding value in the expected data table is found and squared. When all the squared values are summed, the result is the error sum-of-squares (SS) for the fit between the predicted ratings generated from the current set of "guesses" for the SAP's and RRP's and those ratings actually observed. STEPIT is used to successively alter (step) the guesses for the parameters and evaluate the impact on the resulting fit. When changes to the parameter values no longer produce appreciable improvement in the fit (reduction in the error sum-of-squares) between the observed and predicted, MERLIN outputs a series of reports. These reports include the least-squares estimates of the RRP's and SAP's, the complete table of predicted ratings, measures of final fit (multiple R and SS), results of an F-test on the contrast between the fit of the proposed model and the null hypothesis, and so forth. This process requires that one parameter be fixed (i.e., held at a constant value throughout the estimation process) to anchor the scale.

Models Evaluated

Estimates for each of four different statistical models were determined on each cohort's data separately. Model T (theory) was that model derived from the proposed theory. It had one free (RRP) parameter per rater (except for one which was fixed at 500 to anchor the scale) and one free (SAP) parameter per student. Model T permitted, but did not require that, both rater stringency and subject ability contributed curvilinearly to the fit between the predicted and observed ratings. Model T was the most general model considered. Models S and A were derived by imposing restrictions upon Model T.

Model A (ability) imposed the restriction that all raters are equally stringent (all RRP's equal 500) while allowing SAP's to vary. In Model A all variation in ratings is a linear function of the variation in subject ability. This restriction forces each predicted rating for a single student to be the unweighted mean of the observed ratings on this student.

This model corresponds to the common practice of using the mean of the observed ratings as the best measure of the student's ability. Note however that it is accurate only within the context of equal rater stringency or when all raters rate all subjects. When contrasted with Model 0 (null hypothesis), Model A provided a mechanism for determining how well variation in student ability could account for observed ratings. Statistical contrast of Model A (ability) with Model T (both ability and stringency), provides a way to determine if rater stringency contributed to observed ratings beyond that accounted for by student ability. A statistical difference between Models T and A indicates a "stringency main effect".

Model S (stringency) imposed the restriction that all students had equal ability (all SAP's equal 500) while permitting all the RRP's to vary. Thus in Model S variation in ratings is a linear function of variation in rater stringency. When contrasted with Model 0 (null hypothesis), Model S provided a way to determine the extent to which variation in the observed ratings may be accounted for by variation in rater stringency. When contrasted with Model T (i.e., both achievment and stringency), Model S (stringency) provides a way to determine if student ability makes a significant contribution to observed ratings beyond that which could be ascribed to variations in rater stringency. A statistical difference between Models T and S indicates an "ability main effect".

Model 0 embodies the null hypothesis, i.e., a model which accounts for the observed data as chance (random) variation from the overall mean rating (across all raters and students). The null hypothesis error sum of squares (SS) for a particular cohort was equal to the variance of the observed ratings times one less than the number of observations.

Models S and A were not "straw-men" intended to make the proposed model (T) look good. All three hypothetical models were helpful in determining the relationships of interest. Statistical contrasts were done using the F-test:

$$F = ((RME - FME) / df1) / (FME / df2)$$

where:

FME = Full model (FM) error sum-of-squares;
RME = Restricted model (RM) error sum-of-squares;
df1 = free parameters in FM minus free parameters in RM;
df2 = total number of ratings minus free parameters in FM;

and, when the null hypothesis (Model 0) is the restricted model RME is the error sum-of-squares associated with the vector defined by the product of the unit vector (UV) and grand mean.

References

Cason, G.J. MERLIN: A FORTRAN program for finding least-squares estimates of rater reference points, subject ability points, and goodness-of-fit for Cason and Cason's model of performance rating. Little Rock, Arkansas: University of Arkansas for Medical Sciences, 1980.

Chandler, J.P. STEPIT: A FORTRAN II subroutine for finding local minima of real functions. Bloomington, Indiana: Quantum Chemistry Program Exchange, Indiana University, 1965.

Hambleton, R.K., Swaninathan, H., Cook, L.L., Eignor, D.R., & Gifford, J.A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Samejima, F. Homogeneous case of the continuous response model. Psychometrika, 1973, 38(2), 203-219.

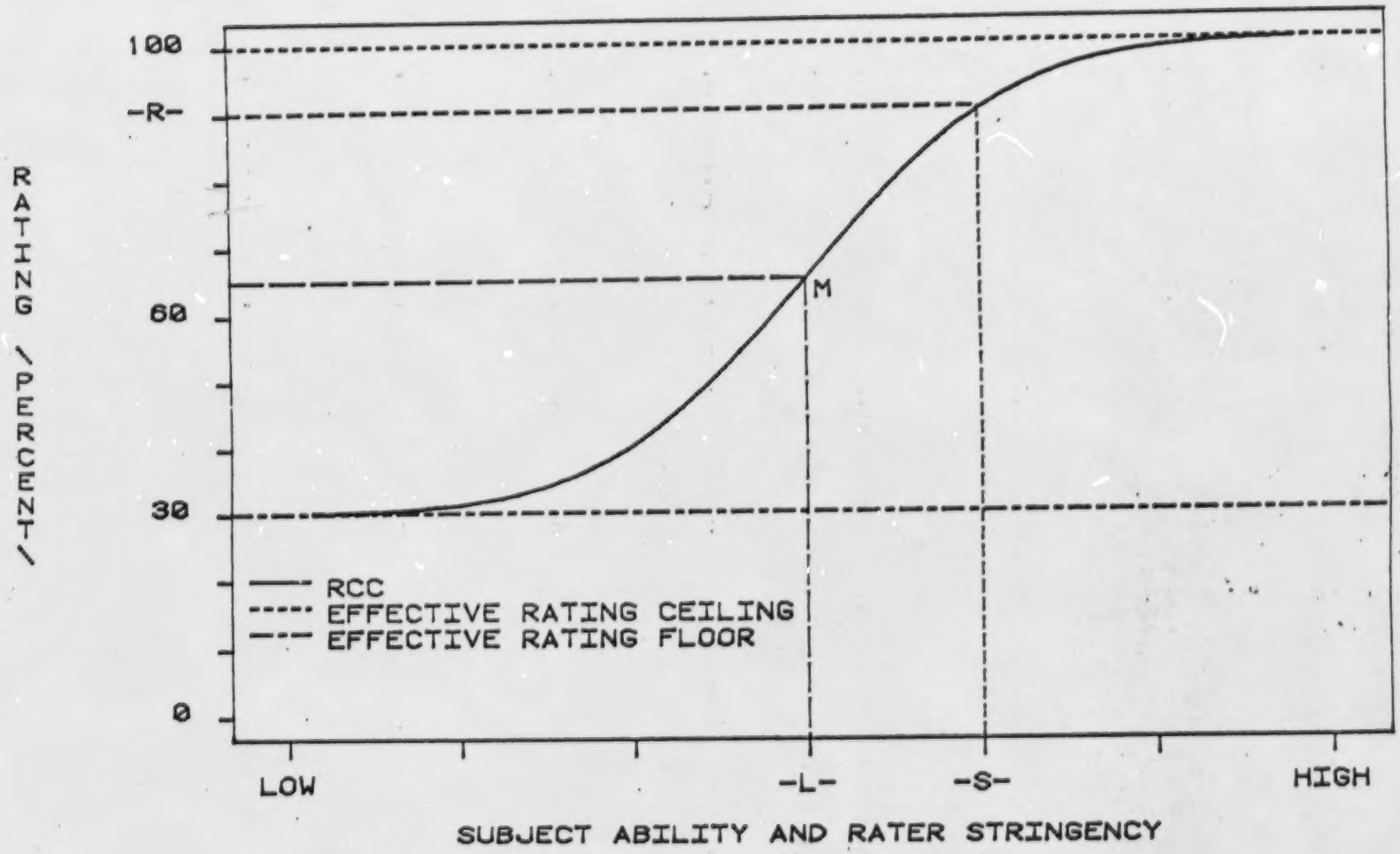

Figure 1. Example rater characteristic curve (RCC): Stringency L gives rating R to ability S.

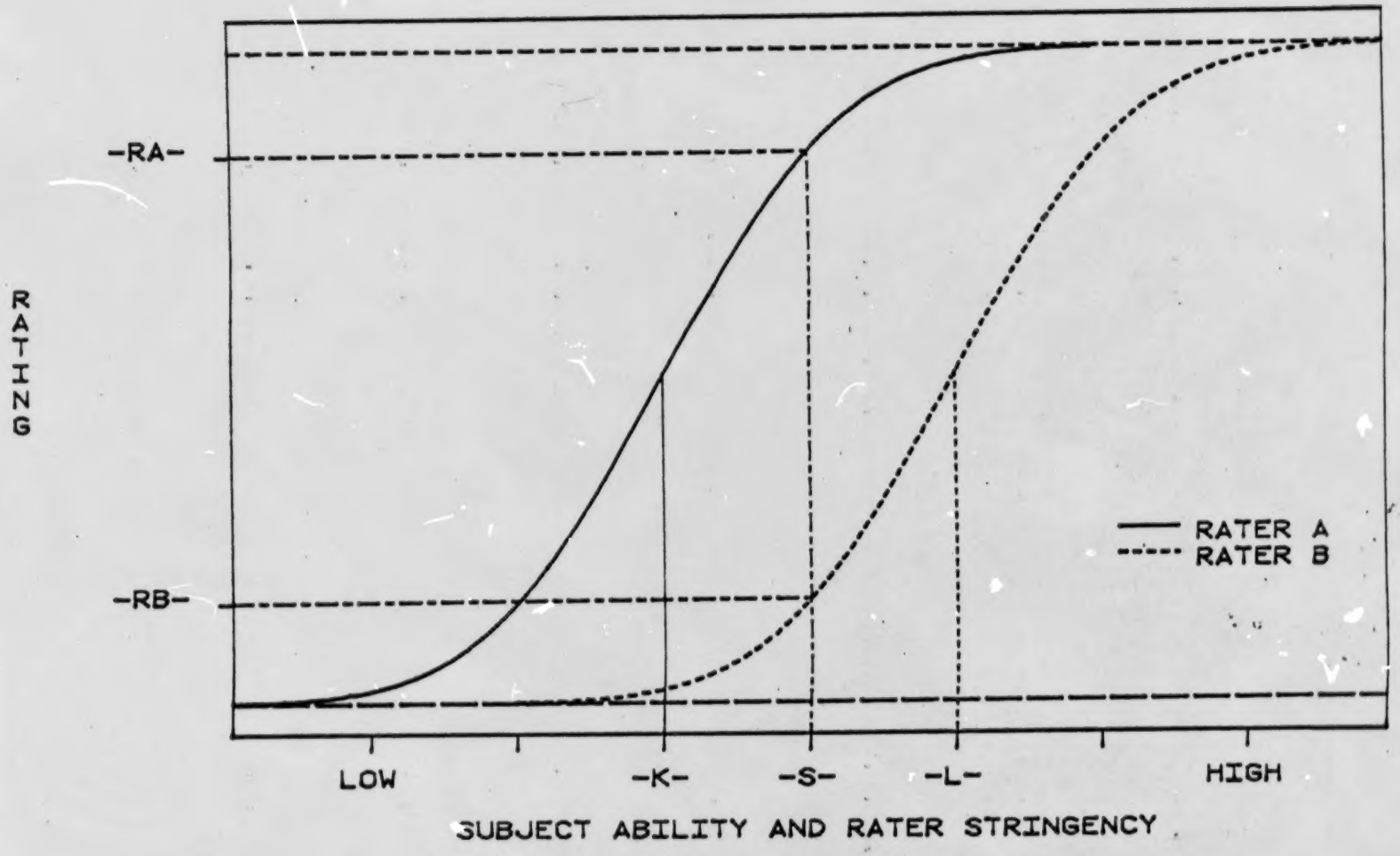

Figure 2. Raters A and B of stringencies K and L give subject of ability S ratings RA and RB respectively.

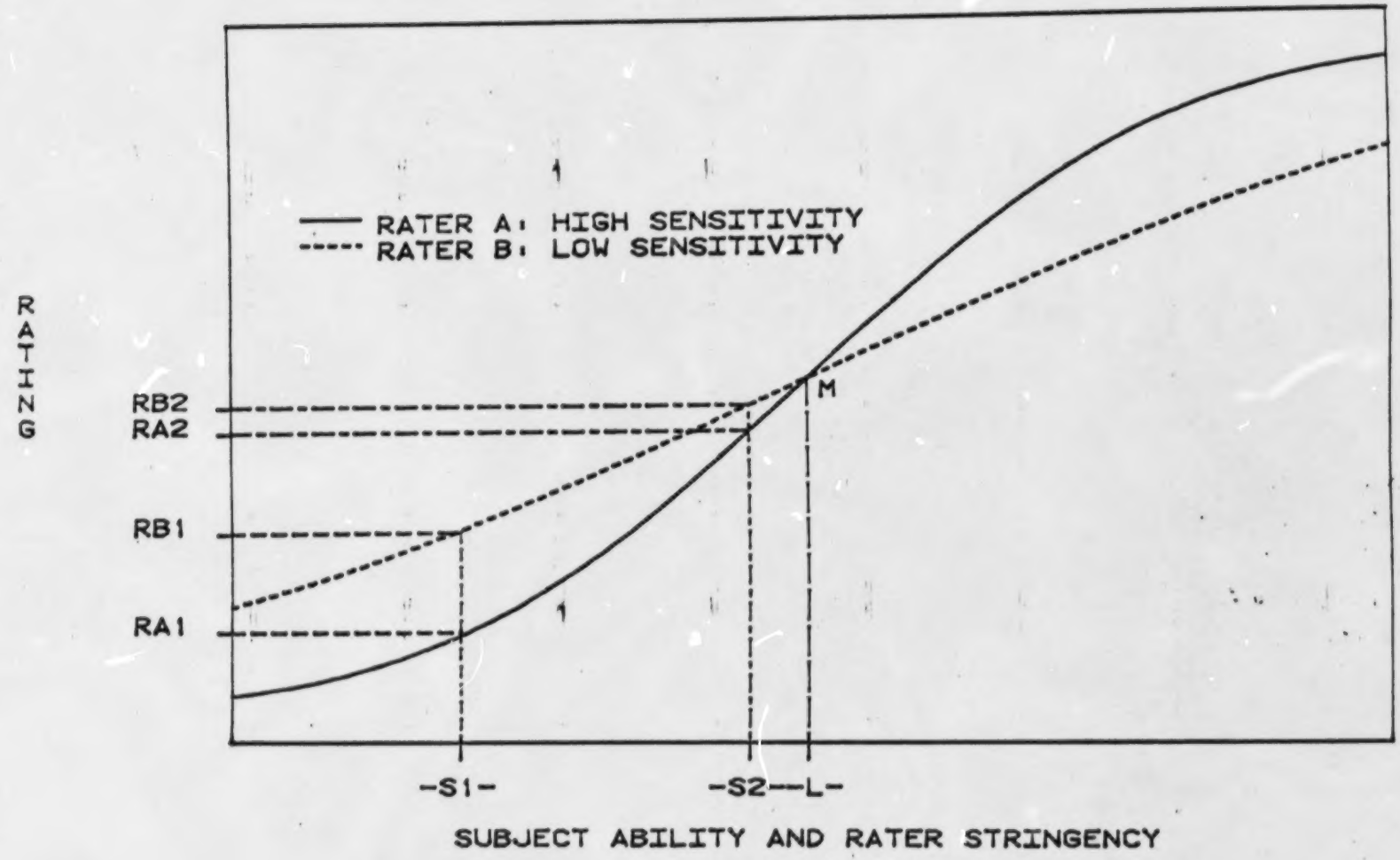

Figure 3. High sensitivity of rater A produces larger difference between ratings (RA1 and RA2) assigned to low ability subject (S1) and high ability subject (S2) than corresponding ratings (RB1 and RB2) assigned by low sensitivity rater B.

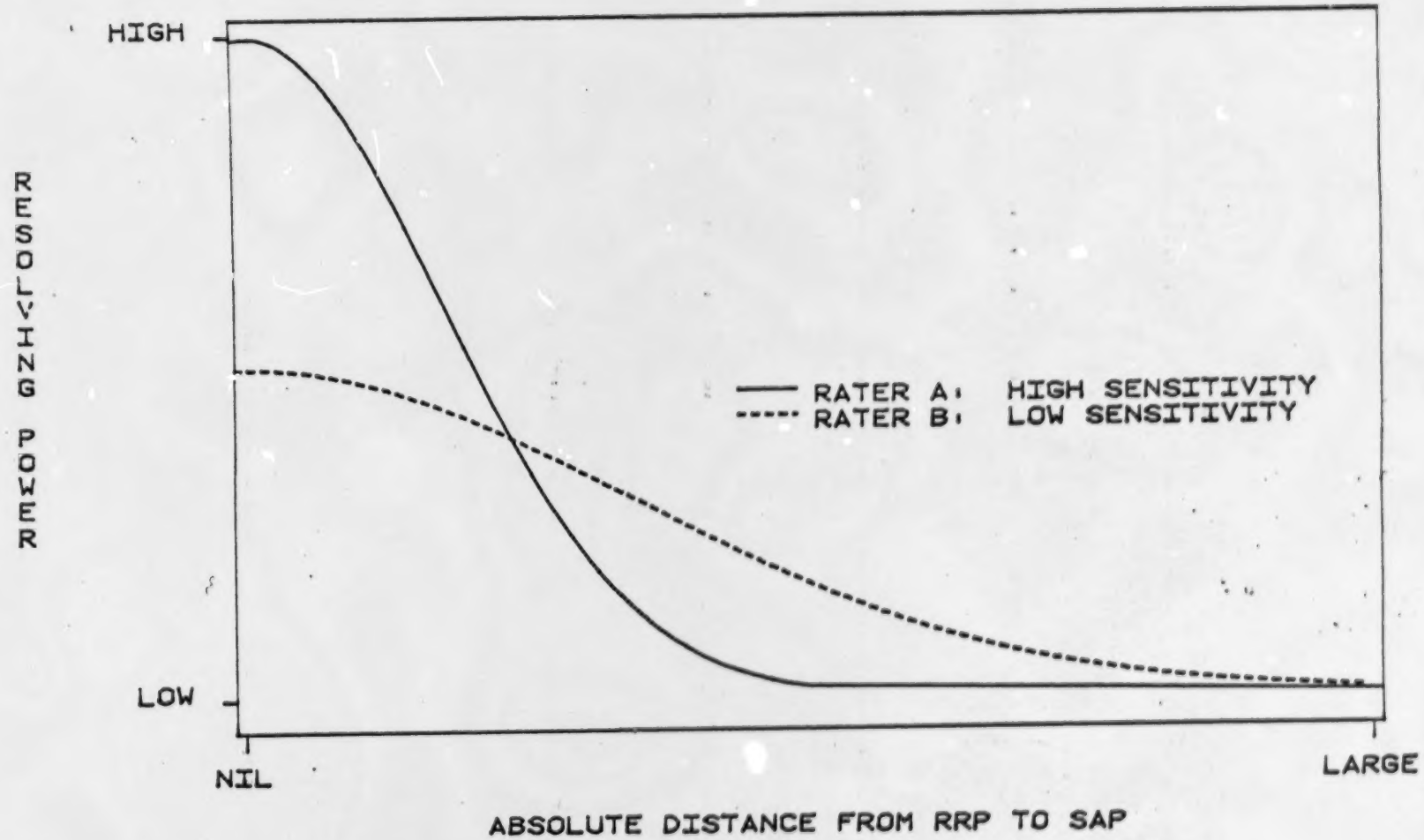

Figure 4. Resolving power as a function of absolute distance from rater reference point (RRP) to subject ability point (SAP) for raters of high and low sensitivity.

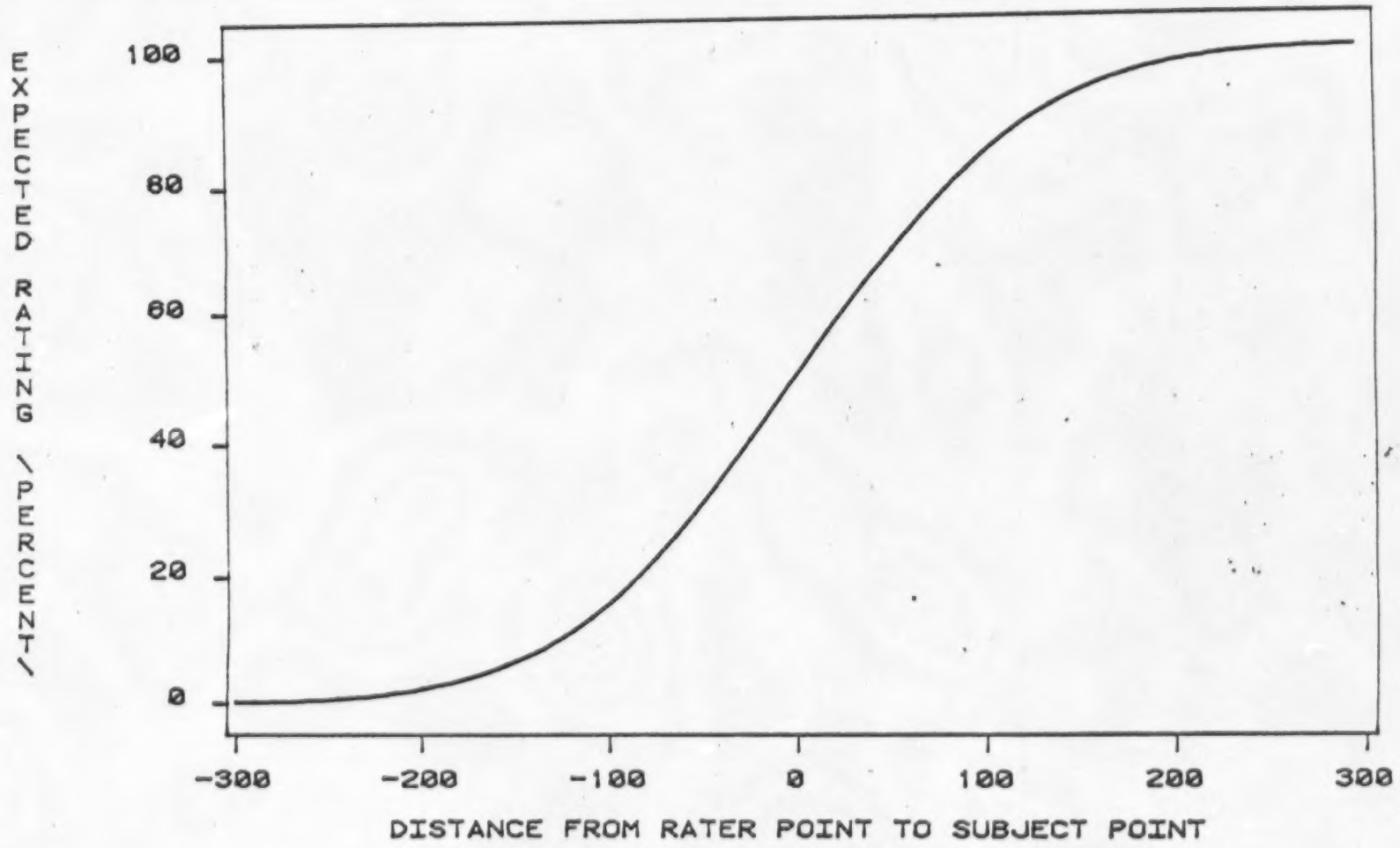

Figure 5. Expected rating as a function of distance between rater reference point and subject ability point.